FORSCHUNG KOMPAKT

Fraunhofer auf der Hannover Messe 2023

Mit Prüftools KI-Systeme vertrauenswürdig und transparent gestalten

ChatGPT hat einen neuen Hype um Künstliche Intelligenz ausgelöst, die Möglichkeiten der KI sind beeindruckend. Gleichzeitig wird die Qualitätssicherung und Kontrolle von KI-Systemen immer wichtiger – insbesondere wenn sie verantwortungsvolle Aufgaben übernehmen. Denn die Ergebnisse des Chatbots beruhen auf riesigen Datenmengen an Texten aus dem Internet. Dabei berechnen Systeme wie ChatGPT allerdings nur die wahrscheinlichste Antwort auf eine Frage und geben diese als Fakt aus. Forschende des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme IAIS präsentieren vom 17. bis 21. April auf der Hannover Messe 2023 am Fraunhofer-Gemeinschaftsstand in Halle 16, Stand A12 verschiedene Prüftools und Verfahren, mit denen sich KI-Systeme systematisch entlang ihres Lebenszyklus auf Schwachstellen untersuchen und gegen KI-Risiken absichern lassen. Die Tools unterstützen Entwickler und Entwicklerinnen sowie Prüfinstitute dabei, die Qualität von KI-Systemen systematisch zu evaluieren und so ihre Vertrauenswürdigkeit sicherzustellen.

Die mediale Omnipräsenz der neuen KI-Anwendung ChatGPT von OpenAI zeigt: Künstliche Intelligenz hat eine beeindruckende Reife erreicht. Der Chatbot, der mit Daten und Texten aus dem ganzen Internet trainiert wurde, reagiert auf Fragen mit Antworten, die sich von von Menschen erstellten Texten nur schwer bis gar nicht unterscheiden lassen. Das macht das KI-System interessant für den vielfältigen Einsatz in Unternehmen – vom Marketing über die Automatisierung der Bearbeitung von Kundenanfragen bis hin zur Generierung von Medieninhalten.

Prüftools für den Blick in die Black Box

In der öffentlichen Diskussion wird allerdings auch Vorsicht angemahnt. Die Kritik richtet sich unter anderem gegen die fehlende Transparenz, etwa aus welchen Quellen der Chatbot seine Antworten generiert. Insbesondere basieren die Vorhersagen auf der Qualität der Input-Daten. »Dies zeigt, wie wichtig es ist, die Güte von KI-Anwendungen systematisch prüfen zu können. Dies gilt vor allem in sensiblen Anwendungsfeldern wie etwa der medizinischen Diagnostik, dem HR-Management, dem Finanzwesen, der Justiz oder in sicherheitskritischen Bereichen, wo KI-Systeme absolut zuverlässige Ergebnisse liefern müssen. Der AI Act – der Europäische Entwurf zur Regulierung von KI-Systemen – stuft diese Beispiele in die Hochrisiko-Kategorie ein und sieht für sie Prüfungen

Kontakt
Roman Möhlmann | Fraunhofer-Gesellschaft, München | Kommunikation | Telefon +49 89 1205-1333 | presse@zv.fraunhofer.de
Silke Loh | Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS | Telefon +49 2241 14-2829 | Schloss Birlinghoven 1 | 53757 Sankt Augustin | www.iais.fraunhofer.de | silke.loh@iais.fraunhofer.de

sogar verpflichtend vor«, sagt Dr. Maximilian Poretschkin, Leiter KI-Absicherung und -Zertifizierung am Fraunhofer IAIS in Sankt Augustin. »Unternehmen, die Hochrisiko-KI-Anwendungen entwickeln oder einsetzen, müssen sich spätestens jetzt dringend mit der Qualitätssicherung ihrer Anwendungen auseinandersetzen.«

Gemeinsam mit seinem Team entwickelt er Prüfwerkzeuge und Verfahren, die KI-Anwendungen in Bezug auf ihre Verlässlichkeit, Fairness, Robustheit, Transparenz oder Datenschutz hin untersuchen und bewerten. Die Tools sind modular miteinander kombinierbar und in ein Software-Framework eingebettet. Die Entwicklung von prototypischen Prüfwerkzeugen wird unter anderem durch das Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen im Rahmen des NRW-Flagship-Projekts ZERTIFIZIERTE KI gefördert. Die zugrunde liegenden Prüfkriterien basieren auf dem KI-Prüfkatalog (siehe Kasten), einem strukturierten Leitfaden für die Praxis, den die Forschenden des Fraunhofer IAIS 2021 veröffentlicht haben.

Neuronale Netze auf ihre Schwachstellen untersuchen

Der Bedarf solcher Prüfwerkzeuge ergibt sich daraus, dass sich KI-Anwendungen oft deutlich von herkömmlicher Software unterscheiden. Letztere ist regelbasiert programmiert, was ein systematisches Durchtesten ihrer Funktionalität erlaubt – also ob die Antworten bzw. Ausgaben in Abhängigkeit der Eingaben korrekt sind. Dies funktioniert bei KI-Anwendungen nicht ohne Weiteres, insbesondere wenn sie auf neuronalen Netzen basieren. Das Werkzeug »ScrutinAI« des Fraunhofer IAIS befähigt Prüferinnen und Prüfer, systematisch nach Schwachstellen von neuronalen Netzen zu suchen und somit die Qualität der KI-Anwendungen zu testen. Ein konkretes Beispiel ist eine KI-Anwendung, die Anomalien und Krankheiten auf CT-Bildern erkennt. Hier stellt sich die Frage, ob alle Arten von Anomalien gleichermaßen gut erkannt werden oder einige besser und andere schlechter. Diese Analyse hilft Prüferinnen und Prüfern zu beurteilen, ob eine KI-Anwendung für ihren vorgesehenen Einsatzkontext geeignet ist. Gleichzeitig können auch Entwicklerinnen und Entwickler profitieren, indem sie Unzulänglichkeiten ihrer KI-Systeme frühzeitig erkennen und entsprechende Verbesserungsmaßnahmen ergreifen können, wie etwa die Anreicherung der Trainingsdaten um spezifische Beispiele.

Der Einsatz des Werkzeugs ist dabei für viele Use Cases denkbar. Das obige Beispiel kann genauso gut durch eine KI-Anwendung ersetzt werden, die Schwachstellen und Materialfehler in sicherheitskritischen Bauteilen detektiert. Auch hier gilt es herauszufinden, ob alle Schwachstellen gleichermaßen gut erkannt werden oder ob es Bereiche in der vorgesehenen Einsatzdomäne gibt, für welche die Leistungsfähigkeit der KI-Anwendung unzureichend ist. »Es geht immer darum, Unzulänglichkeiten im neuronalen Netz zu erkennen, wenn auch in unterschiedlichen Kontexten«, erläutert Poretschkin.

Unsicherheiten einschätzen

Die in das Framework integrierte und vom Fraunhofer IAIS entwickelte Methode »uncertAInty« stattet neuronale Netze mit einer situationsabhängigen Güteeinschätzung aus, mit der diese ihre eigene Sicherheit bezüglich der gemachten Vorhersage bewerten. »Bei hochautomatisierten KI-Entscheidungen ist es wichtig beurteilen zu können, wie sicher sich eine KI mit ihrem Ergebnis ist. Konkret muss etwa ein autonomes Fahrzeug Objekte und Menschen in seiner Umgebung zuverlässig erkennen können, damit es angemessen darauf reagieren kann. Die Unsicherheitsbewertung hilft hierbei zu messen, wie stark man der Entscheidung des Systems vertrauen kann oder ob bestimmte Fallback-Mechanismen aktiviert werden müssen bzw. ein Mensch die finale Entscheidung treffen muss«, so Poretschkin. Die uncertAInty-Methode stellt somit einen wichtigen Baustein zur Absicherung von KI-Anwendungen dar, um sie für sensible Einsatzkontexte nutzbar zu machen.

Vergleich von KI-Modellen

Mit dem »Benchmarking«-Werkzeug lässt sich schließlich untersuchen, welches KI-Modell sich am besten für eine bestimmte Aufgabe eignet. »Es gibt eine Flut neuer KI-Anwendungen, die Unternehmen in ihre Prozesse integrieren können. Benchmarking hilft bei der geeigneten Auswahl«, sagt der Forscher. Das Tool umfasst unter anderem auch eine Funktionalität, um die Fairness von Trainingsdatensätzen messen zu können. Dies ist beispielsweise in der HR-Branche entscheidend, wenn es um KI-Anwendungen geht, die bei der Auswahl von neuen Mitarbeitenden unterstützen. Hier muss die KI-Anwendung mit ausgewogenen und statistisch repräsentativen Datensätzen trainiert werden, um eine Benachteiligung von Personengruppen zu vermeiden und die Chancengleichheit zu gewährleisten

Auf der Hannover Messe am Fraunhofer-Gemeinschaftsstand in Halle 16, Stand A12 demonstriert das Team des Fraunhofer IAIS anhand eines interaktiven Demonstrators aus dem medizinischen Bereich, wie eine KI-Anwendung mithilfe der Prüfwerkzeuge systematisch auf Qualität untersucht werden kann. Darüber hinaus erfahren Interessierte, wie eine Prüfung von KI in Unternehmen konkret erfolgen kann.

KI-Prüfkatalog

Von Sprachassistenzsystemen über die Analyse von Bewerbungsdokumenten bis hin zum autonomen Fahren – als Schlüsseltechnologie der Zukunft kommt Künstliche Intelligenz (KI) überall zum Einsatz. Umso wichtiger ist es, KI-Anwendungen so zu gestalten, dass sie verlässlich und sicher agieren und transparent und zuverlässig mit Daten umgehen. Dies ist eine notwendige Voraussetzung dafür, dass KI auch in sensiblen Bereichen zum Einsatz kommen kann und Nutzer sowie Nutzerinnen nachhaltig Vertrauen in die Technologie haben. Mit dem KI-Prüfkatalog des Fraunhofer IAIS erhalten Unternehmen einen praxisorientierten Leitfaden, der sie befähigt, ihre KI-Systeme vertrauenswürdig zu gestalten. Auf rund 160 Seiten beschreibt er, wie KI-Anwendungen systematisch hinsichtlich Risiken evaluiert werden können, formuliert Vorschläge für Prüfkriterien zur Messung der Qualität der Systeme und schlägt Maßnahmen vor, die KI-Risiken mindern können. Der KI-Prüfkatalog wurde vom Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen gefördert. Er steht kostenlos zum Download zur Verfügung: KI-Prüfkatalog - Fraunhofer IAIS

Abb. 1 Das Tool ScrutinAI ermöglicht es, Fehler in KI-Modellen oder Trainingsdaten aufzudecken und die Ursachen dafür zu analysieren. Im vorliegenden Beispiel wird ein KI-Modell zur Erkennung von Anomalien und Krankheiten auf CT-Bildern untersucht.

© Fraunhofer IAIS

Abb. 2 Mit dem KI-Prüfkatalog des Fraunhofer IAIS erhalten Unternehmen einen praxisorientierten Leitfaden, der sie befähigt, ihre KI-Systeme vertrauenswürdig zu gestalten.

© Fraunhofer IAIS